Tome 201 — Septembre-octobre-novembre-décembre — N^os 7-8-9 2017

BULLETIN
DE
L'ACADÉMIE NATIONALE DE MÉDECINE

publié par

MM. Daniel Couturier, Secrétaire perpétuel
et Jean François Allilaire, Secrétaire adjoint

Rédacteur en chef : Professeur Jean-Noël Fiessinger
Adjointe à la Rédaction : Sibylle du Chaffaut

ACADÉMIE NATIONALE DE MÉDECINE
16, RUE BONAPARTE — 75272 PARIS CEDEX 06
http://www.academie-medecine.fr

Bull. Acad. Natle Méd., 2017, *201*, nos 7-8-9, 1197-1201, séance du 7 novembre 2017

COMMUNICATION

Système de veille automatique pour la détection de maladies animales émergentes

Mots-clés : Veille épidémiologique. Fouille de Textes.

An automatic surveillance system for emerging animal disease detection

Key-words: Épidémiologic Surveillance. Text Mining.

Mathieu ROCHE *,**

RÉSUMÉ

Cet exposé présente une méthode automatique d'extraction d'informations sur les maladies animales à partir du Web (PADI-web). PADI-web est un outil de fouille de textes pour la détection automatique, la catégorisation et l'extraction d'informations liées aux épidémies à partir d'articles de presse issus du Web. PADI-web se concentre actuellement sur cinq maladies animales exotiques et infectieuses et huit syndromes chez cinq animaux hôtes.

SUMMARY

This talk deals with a an automatic method for automatic extraction of animal disease information from the Web (PADI-web). PADI-web is a text mining tool for automatic detection, categorization, and extraction of disease outbreak information from Web news articles. Currently PADI-web monitors the Web for five exotic animal infectious diseases and eight syndromes related to five animal hosts.

* Cirad, TETIS, Montpellier, France
** TETIS, Univ. Montpellier, APT, Cirad, CNRS, Irstea, Montpellier, France

Tirés à part : M. Mathieu Roche, UMR TETIS – MTD, 500 rue J.F. Breton, 34093 Montpellier Cedex 5.
Article reçu le 21 octobre 2017, accepté le 6 novembre 2017

INTRODUCTION ET CONTEXTE

Dans le contexte de la « Veille sanitaire internationale » de la Plateforme nationale d'epidemiosurveillance en santé animale (Plateforme ESA), le Cirad, l'ANSES et la Direction générale de l'alimentation (DGAl) développent, depuis 2013, un système de veille automatique du Web qui effectue :

— La collecte quotidienne de dépêches épidémiologiques provenant de sources non officielles incluant les médias électroniques.
— L'extraction automatique d'informations issues de ces dépêches.
— Une restitution synthétique et agrégée de l'information (cartes, séries spatiotemporelles, etc.).

Le processus automatique de veille sanitaire en santé animale sur le Web [1] fondée sur les évènements est décliné en 4 principales étapes :

1) La collecte automatique d'articles sanitaires via des requêtes Web (sur la base de mots-clés).
2) La classification automatique des articles collectés selon leur contenu : *pertinents* (articles qui décrivent des évènements sanitaires liés à l'apparition des foyers de maladies exotiques ou nouvelles) et *non pertinents* (tout autre article).
3) L'extraction automatique de l'information sanitaire à partir des articles pertinents (maladie, date et lieu d'évènement, signes cliniques, hôtes touchés, etc.).
4) L'analyse et l'évaluation du processus à l'aide de connaissances d'experts du domaine.

Les maladies actuellement « surveillées » par ce système de veille sont *la peste porcine africaine*, *la grippe aviaire*, *la fièvre catarrhale ovine*, *la fièvre aphteuse* et *la maladie de Schmallenberg*. Le système est développé dans un cadre générique et permet la veille d'autres maladies.

Les sections suivantes synthétisent les approches automatiques mises en œuvre pour réaliser les trois premières étapes du processus.

ÉTAPE 1 : COLLECTE DES DONNÉES

Pour collecter des articles en provenance de sources non officielles du Web (dépêches, articles de journaux locaux, etc.), nous proposons une approche fondée sur des requêtes Web composées de termes tels que :

— des *noms de maladies* pour la ***veille propre à des maladies connues***,
— des combinaisons entre *des signes cliniques* et *des hôtes* pour une ***veille syndromique***.

Pour **identifier les termes utiles pour la veille**, nous avons effectué deux types d'actions qui se sont révélées parfaitement complémentaires :

— Nous avons recueilli directement, auprès des experts, les mots-clés qui caractérisent les maladies.

— Nous avons constitué un corpus (ensemble de données textuelles composé d'articles scientifiques et articles du Web) et effectué une tâche de *fouille de textes* sur celui-ci permettant d'obtenir des mots-clés caractéristiques qui ont été validés par les experts.

Pour l'identification des termes par fouille de textes, nous utilisons BioTex développé dans le cadre de l'ANR SIFR [1] (Semantic Indexing of French Biomedical Data Resources project). BioTex [2] exploite à la fois des informations statistiques et linguistiques pour extraire une terminologie à partir de textes libres. Les termes candidats sont tout d'abord retenus s'ils respectent des patrons syntaxiques définis (nom-adjectif, nom-préposition-nom, etc.) qui peuvent être spécifiques au domaine biomédical. Après un tel filtrage linguistique, un filtrage statistique est appliqué. Celui-ci mesure l'association entre les mots composant un terme en utilisant une mesure appelée C-Value [3] et en intégrant une pondération (TF-IDF — Term Frequency — Inverse Document Frequency). Le but de C-Value est d'améliorer l'extraction des termes complexes (composés de plusieurs mots) alors que la pondération TF-IDF met en exergue le pouvoir discriminant du terme candidat.

ÉTAPE 2 : CLASSIFICATION

Le but de cette étape est d'**identifier les documents pertinents** qui seront exploités à l'étape suivante. Le processus proposé et en cours d'évaluation s'appuie sur des méthodes de classification qui consistent à apprendre automatiquement un modèle à partir d'un corpus d'articles manuellement étiquetés en trois catégories : « nouveaux cas », « bilan » et « général ». Dans ce processus d'apprentissage, le corpus doit alors être au préalable représenté sous un format exploitable par les algorithmes ; dans notre cas nous avons utilisé le modèle de Salton [4]. Ce dernier consiste à représenter un corpus par une matrice où les lignes sont relatives aux descripteurs linguistiques (en général, *les mots*) et les colonnes sont associées aux documents. Les cellules de ces matrices donnent la fréquence d'apparition d'un descripteur dans un document. Ainsi, la matrice formée peut être utilisée pour effectuer diverses tâches automatiques, en particulier l'identification des documents pertinents (tâche de classification par apprentissage automatique) qui seront exploités à l'étape suivante (étape 3) du processus.

[1] http://sifr.strikingly.com

ÉTAPE 3 : EXTRACTION D'INFORMATION

L'extraction d'information (EI) est une tâche qui consiste à identifier automatiquement un ensemble d'informations jugées pertinentes à partir de données textuelles [5]. En nous appuyant sur ce principe, nous proposons une approche d'extraction automatique d'information à partir des documents collectés (étape 1) et classifiés (étape 2). Un des objectifs consiste à standardiser l'information sanitaire extraite à partir de données non-structurées issues du Web selon les mêmes principes que les données issues des sources officielles (OIE [2], FAO [3], ADNS [4]). Ceci pourra finalement produire un ensemble de données comparables et adaptées pour une analyse épidémiologique complète des événements sanitaires au niveau international.

L'approche de fouille de textes proposée permet l'**extraction de diverses informations épidémiologiques**, à savoir les *maladies*, les *hôtes*, les *signes cliniques*, les *lieux*, les *dates*, les *nombres de cas*. Un événement sanitaire doit être associé à un *lieu* (représenté par des coordonnées géographiques, une zone administrative, un pays), une *date*, ainsi qu'une *espèce touchée*, avec soit une *maladie confirmée*, soit des *signes cliniques*. Notre approche d'extraction d'information (EI) par fouille de données repose sur une méthode d'identification automatique de motifs en utilisant différents dictionnaires et traitements [5]. Ensuite, les règles découvertes automatiquement à partir d'un corpus manuellement annoté sont utilisées comme descripteurs dans un modèle d'apprentissage automatique (*machine learning*) pour prédire la pertinence des entités identifiées (lieux, dates, hôtes, etc.).

CONCLUSION

Cette étude apporte plusieurs contributions méthodologiques pour les dispositifs de veille sanitaire internationale en France (VSI). Tout d'abord, nous avons proposé une démarche générale et automatisée dédiée à la veille sur le Web de maladies animales infectieuses et exotiques. Comparativement aux autres systèmes de biosurveillance, notre approche se concentre sur une liste de maladies spécifiques, établie par des experts en santé animale. Avec ce type d'approche, nous pouvons extraire l'information épidémiologique importante à partir de données Web afin de fournir aux utilisateurs une information pertinente en temps réel sur les principales émergences sanitaires au niveau international. Nos résultats préliminaires sont très encourageants ; d'autres études devront être menées pour approfondir les différentes facettes de ce domaine complexe.

---

[2] World Organisation for Animal Health

[3] Food and Agriculture Organization of the United Nations

[4] Animal Disease Notification System

[5] *Geonames* utilisé pour identifier les noms de lieux (pays, régions, villes, villages etc.), *HeidelTime* utilisé pour marquer toutes les dates dans le texte, des dictionnaires de noms de maladies, de signes cliniques et d'hôtes obtenus avec un processus de fouille de textes.

INFORMATIONS COMPLÉMENTAIRES

Contexte :

Cette communication est une synthèse d'un travail mené collectivement avec de nombreux chercheurs et étudiants dans un cadre pluridisciplinaire [1], en particulier durant la thèse d'**Elena Arsevska** encadrée par **Renaud Lancelot**, **Barbara Dufour** et **Mathieu Roche**. Les autres chercheurs et ingénieurs (Cirad, INRA, Université de Montpellier) impliqués dans le projet sont : **Sylvain Falala, Alizé Mercier, Sarah Valentin, Samiha Fadloun, Jocelyn De Goër De Herve, Julien Rabatel, Pascal Poncelet, Arnaud Sallaberry**

Remerciements :

Ce travail est soutenu par la DGAL et par le projet SONGES (Région Occitanie et Fonds Européens de Développement Régional).

Lien :

http://www.cirad.fr/nos-recherches/resultats-de-recherche/2016/veille-sanitaire-sur-le-web-un-outil-pour-prevenir-la-propagation-des-maladies-animales

Contact :

Renaud Lancelot : renaud.lancelot (a) cirad.fr (Cirad, ASTRE)

Mathieu Roche : mathieu.roche (a) cirad.fr (Cirad, TETIS)

RÉFÉRENCES

[1] Arsevska, E., M. Roche, P. Hendrikx, D. Chavernac, S. Falala, R. Lancelot et B. Dufour. 2016. Identification of terms for detecting early signals of emerging infectious disease outbreaks on the web. *Computers and Electronics in Agriculture* 123, p. 104-115.

[2] Lossio Ventura J.A., Jonquet C., Roche M., Teisseire M. 2016. Biomedical term extraction: overview and a new methodology, *Information Retrieval Journal* — Springer, Volume 19, Issue 1, p.59-9.

[3] Frantzi K., Ananiadou S., Mima H. 2000. Automatic recognition of multi-word terms: the C-value/NC-value method. *Int. Jour. on Digital Libraries*, 3(2), p. 115-130.

[4] Salton, G. 1983. Introduction to Modern Information Retrieval. English. New York: Mcgraw-Hill College, p. 440.

[5] Poibeau, T. 2003. Extraction automatique d'information : du texte brut au Web seman- tique. Lavoisier, p. 238.

Printed in France *Le Directeur de la publication* M. Daniel COUTURIER.

Imprimé par l'Imprimerie F. Paillart
86, chaussée Marcadé 80100 Abbeville
Académie de Médecine, éditeur, Paris — Dépôt légal 2e trimestre 2018 — No 0190.